On the Planning Abilities of Large Language Models - A Critical Investigation

Abstract

Intrigued by the claims of emergent reasoning capabilities in LLMs trained on general web corpora, in this paper, we set out to investigate their planning capabilities. We aim to evaluate (1) the effectiveness of LLMs in generating plans autonomously in commonsense planning tasks and (2) the potential of LLMs as a source of heuristic guidance for other agents (AI planners) in their planning tasks. We conduct a systematic study by generating a suite of instances on domains similar to the ones employed in the International Planning Competition and evaluate LLMs in two distinct modes: autonomous and heuristic. Our findings reveal that LLMs' ability to generate executable plans autonomously is rather limited, with the best model (GPT-4) having an average success rate of ~12% across the domains. However, the results in the heuristic mode show more promise. In the heuristic mode, we demonstrate that LLM-generated plans can improve the search process for underlying sound planners and additionally show that external verifiers can help provide feedback on the generated plans and back-prompt the LLM for better plan generation.